Asymmetric Spatio-Temporal Embeddings for Large-Scale Image-to-Video Retrieval

Noa Garcia and George Vogiatzis

Aston University, UK

INTRODUCTION

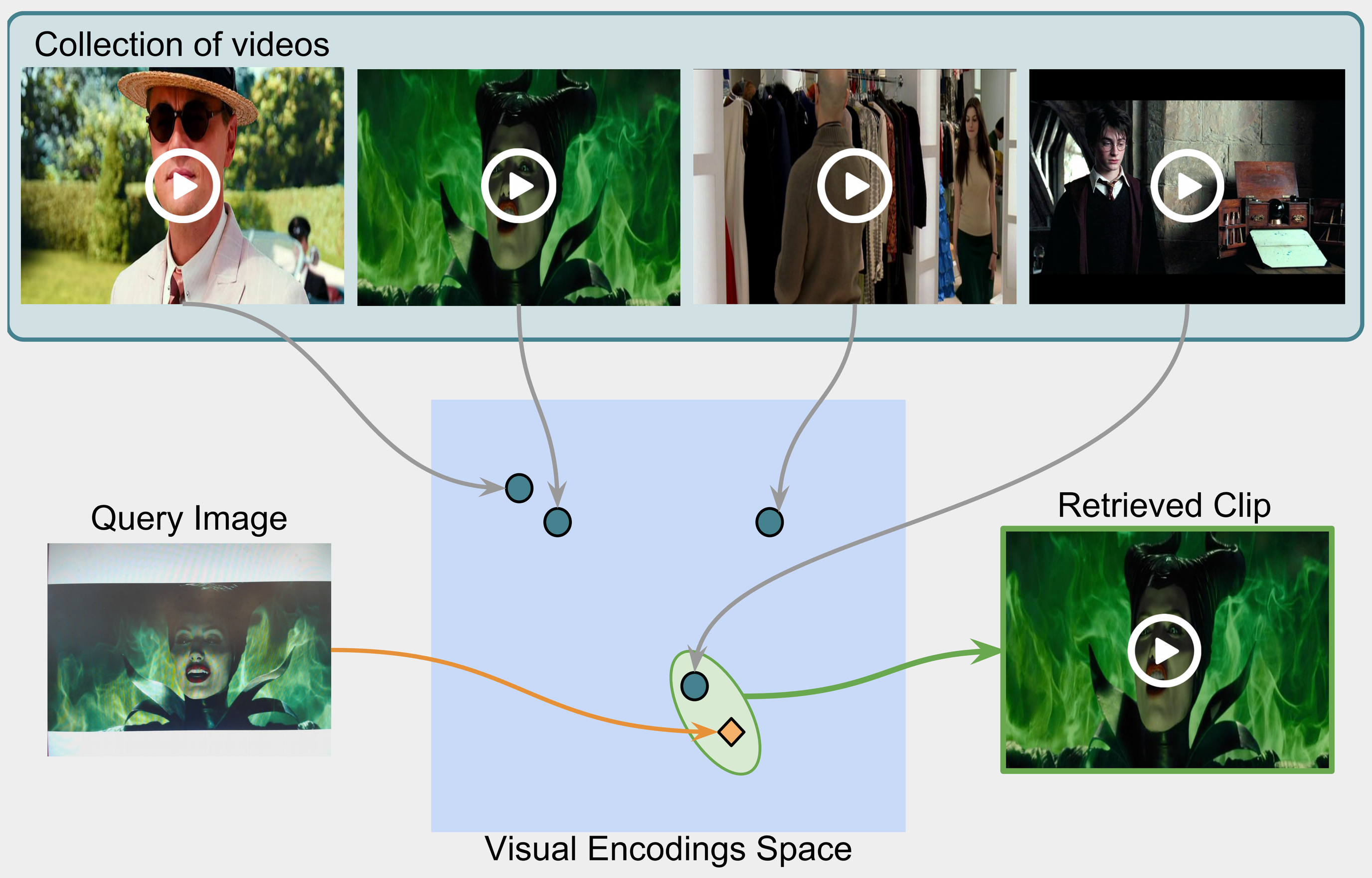


Image-to-Video Retrieval

Finding video clips in a large-scale collections using static images.

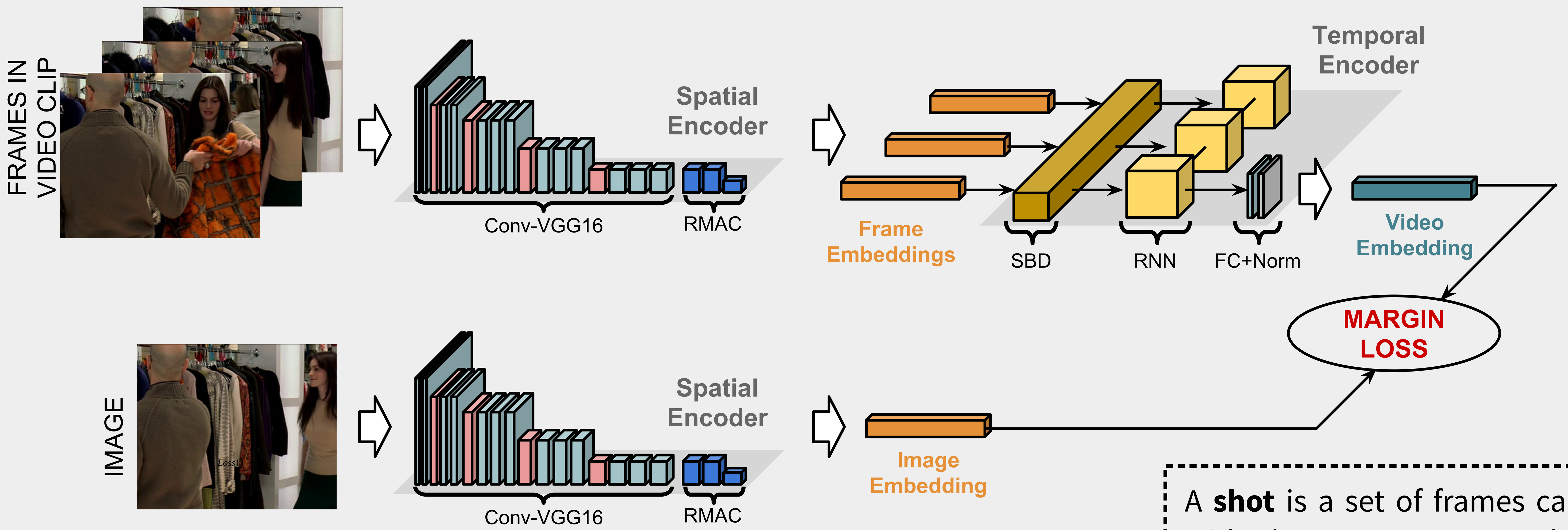
Challenges

- Asymmetry: different processing tools for images and videos.
- Scalability: the number of frames scales very fast.
- Efficiency: to reduce the amount of data to be processed.

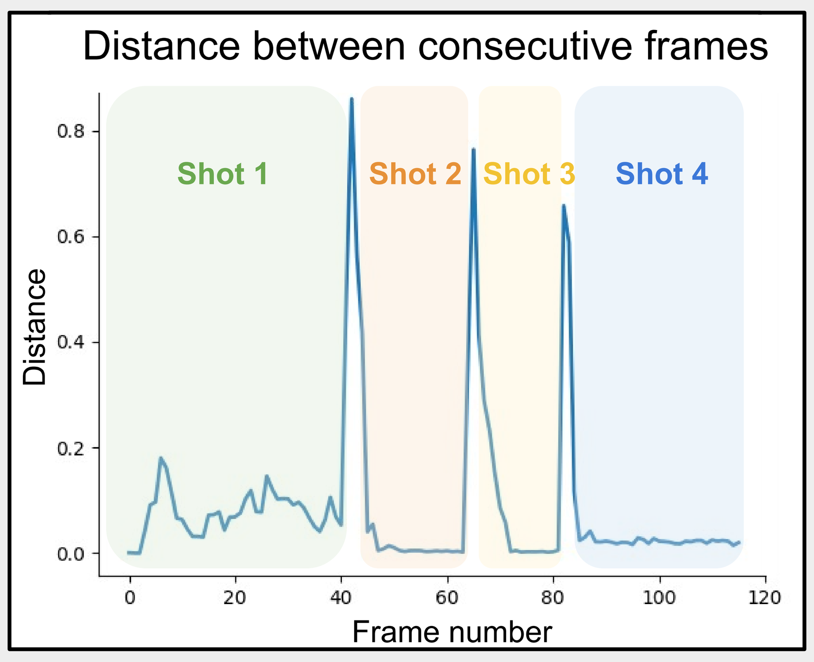
We propose

To encode images and videos into a common embedding space using an asymmetric spatio-temporal encoder.

MODEL



A **shot** is a set of frames captured with the same camera. The **SBD** detects shot boundaries when the distance between frames is large.



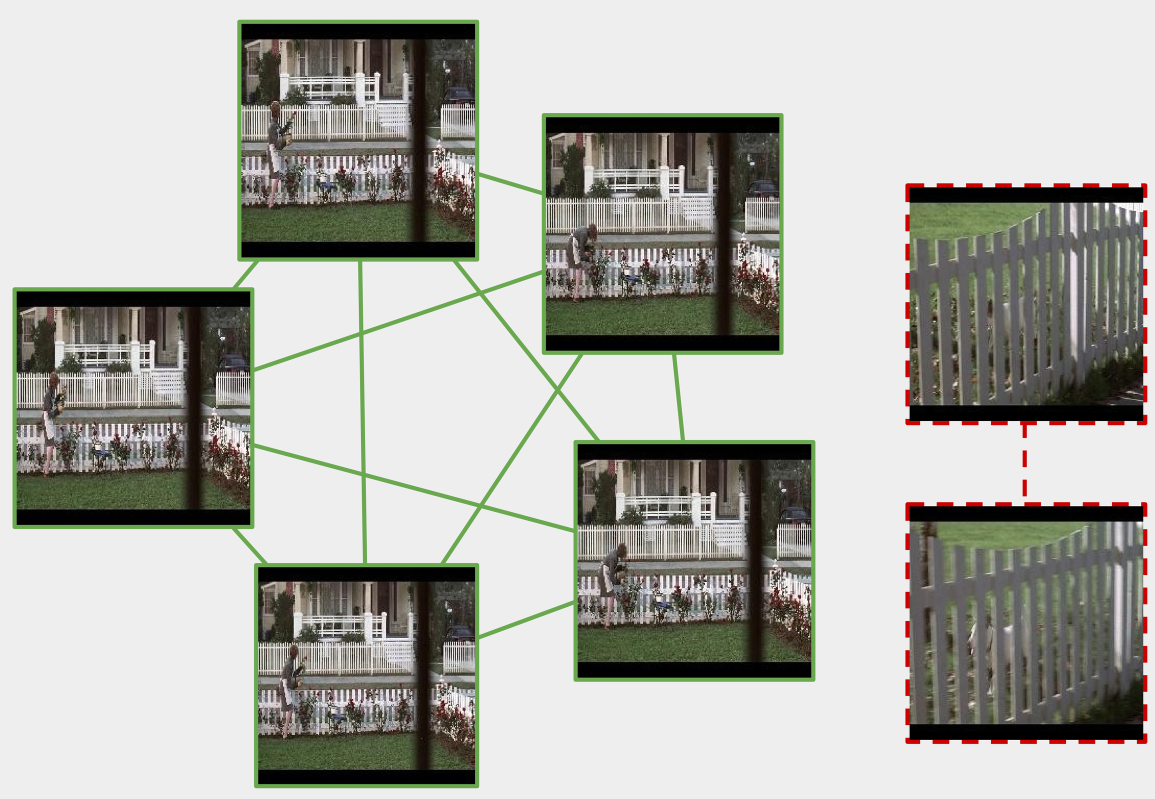
TRAINING

$$\mathrm{Loss}(\mathcal{F}_i, \vartheta_i) = y_i(1 - \cos(\mathcal{F}_i, \vartheta_i)) + (1 - y_i)(\max(0, \cos(\mathcal{F}_i, \vartheta_i) - \Delta))$$

We use pairs of **{frame, video}** for training and we ensure that the distance between **matching** pairs (y = 1) is less than the distance between **non-matching** pairs (y = 0).

Training Data

- LSMDC (Rohrbach et al., 2017) with 40 movies and 26,496 clips.
- Shots obtained from clips using **data graphs**.



RESULTS

Method	dim	SI2V [5]	VB [6]
Scene FV [3]	65,536	0.500	**0.622**
Sum-Pool Alexnet FC6 [3]	4,096	0.071	0.012
Sum-Pool AlexNet FC7 [3]	4,096	0.065	0.013
Sum-Pool VGG16 FC6 [3]	4,096	0.067	0.013
Sum-Pool VGG16 FC7 [3]	4,096	0.069	0.011
Ours (LSTM)	**512**	0.602	0.580
Ours (GRU)	**512**	**0.606**	0.572

[3] Araujo and Girod, 2017, [5] Araujo et al., 2015, [6] Araujo et al., 2016.